Name:_____ Date: _____

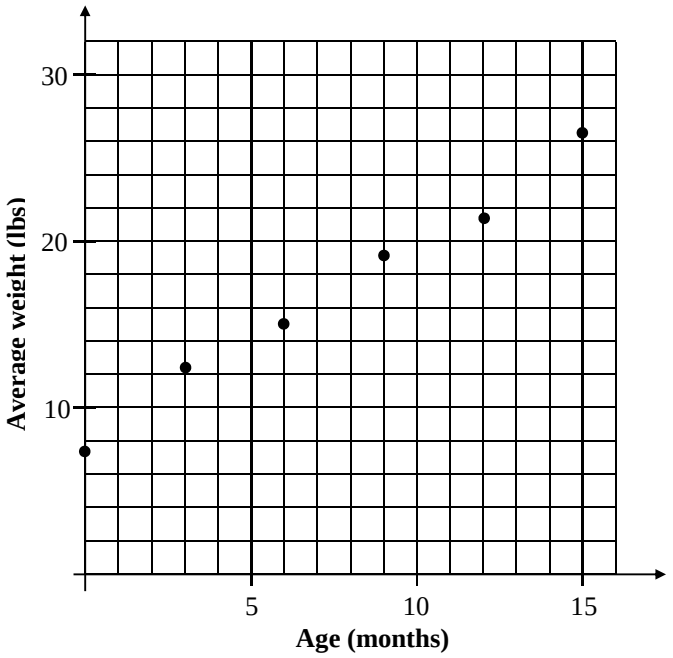
A2CC: Regression

Oftentimes in science, a mathematical relationship between two variables is desired for predictive purposes. In the real world, the relationship between two variables is not always a perfect one, thus we often look for the "best" curve that can fit the data. Today we will review how to do this with a linear function.

***Exercise* #1:** A pediatrician would like to determine the relationship between infant female weights versus age. The pediatrician studies 100 newborn girls and finds their average weight at the end of 3 month intervals. The data is shown below and graphed on the scatter plot.

Age (months)	0	3	6	9	12	15
Average Weight (pounds)	7.2	12.2	15.1	19.4	21.5	26.3

(a) Using a ruler, draw a line that you think best fits this data. As a general guideline, try to draw it such that there are as many data points above the line as below it.



(b) By picking two points that are on the line (not necessarily data points), determine the equation of your best fit line. Round your coefficients to the nearest *tenth*.

(c) Using the linear regression command on your calculator, find the equation of the best fit line for this data. Round all **linear parameters** to the nearest *tenth*.

(d) Use your calculator to determine the **linear correlation coefficient.** Round to the nearest *thousandth*. How can you interpret this value in terms of the variation in weight due to age?

***Exercise* #2:** Using the equation that your calculator produced in Exercise #1, predict the weight of a baby girl after 10 months. Round your answer to the nearest tenth of a pound.

The use of a model to predict outputs when the input is within the range of the known data is called **interpolation**. Interpolation tends to be fairly accurate.

***Exercise* #3:** Using the equation that your calculator produced in Exercise #1, predict the weight of a baby girl after 2 years. Round your answer to the nearest tenth of a pound.

The use of a model to predict outputs when the input is outside of the range of the known input data is called **extrapolation**. Models are most helpful when they can be used to extrapolate, but tend to be less accurate.

***Exercise* #4:** Biologists are trying to create a least-squares regression equation (another name for best fit line) relating the length of steelhead salmon to their weight. Seven salmon were measured and weighed with the data given below.

Length (inches)	22	24	28	34	39	42	48
Weight (pounds)	3.43	4.46	7.08	14.21	22.19	31.22	35.67

(a) Determine the least-squares regression equation, in the form $y = ax + b$, for this data. Round all coefficients to the nearest *hundredth*.

(b) Using your equation from part (a), determine the expected weight of a salmon that is 30 inches long.

(c) Using your equation from part (a), determine the expected weight of a salmon that is 52 inches long.

(d) In which part, (b) or (c), did you use interpolation and in which part did you use extrapolation? Explain.

Just as we fit data with a linear model we can also fit with all sorts of other mathematical models, depending on the context of the situation. In this lesson we will examine **exponential regression** and **sinusoidal regression**. You could be asked to run a quadratic regression, logarithmic regression, power regression, …. The process is similar for each and all are found in Stat Calc menu. Exponential regression is review from Common Core Algebra I, so we will start with that.

***Exercise* #5:** The population of Jamestown has been recorded for selected years since 2000. The table below gives these populations.

Year	2002	2004	2005	2007	2009
Population	5564	6121	6300	6812	7422

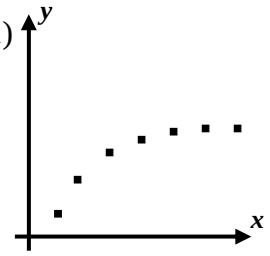
(a) Using your calculator, determine a best fit exponential equation, of the form $y = a \cdot b^x$, where x represents the number of years since 2000 and y represents the population. Round a to the nearest integer and b to the nearest *thousandth*.

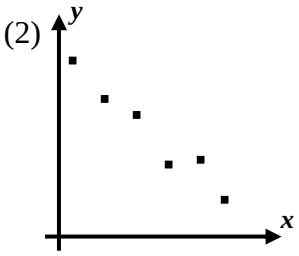
(b) Sketch a graph of the exponential function for the years 2000 to 2050. Label your window and your y-intercept.

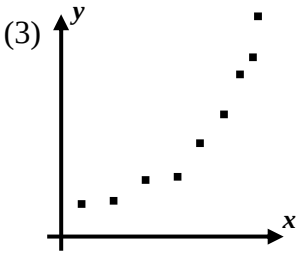
(c) By what percent does your exponential model predict the population is increasing per year? Explain.

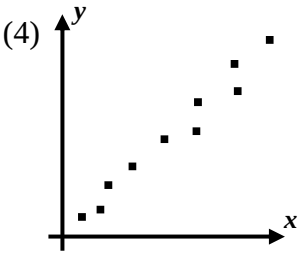
(d) Algebraically determine the number of years, to the nearest year, for the population to reach 20 thousand.

***Exercise* #6:** Which of the following scatter plots would be best fit with an exponential equation?

(1) 

(2) 

(3) 

(4) 

Sinusoidal, or trigonometric, regression is much more complicated than either linear or exponential. It should be used in situations that appear **periodic** in nature.

***Exercise* #7:** The temperature of a chemical reaction changes during the reaction. The temperature was measured every two minutes and the data is shown in the table below.

Time (min)	0	2	4	6	8	10	12	14	16	18	20
Temp $(^\circ\text{C})$	35.7	38.9	41.6	42.3	40.8	38.4	36.1	34.2	35.9	39.1	41

(a) Why does it seem like this data might be periodic? Create a quick scatter plot using your calculator to verify.

(b) Use your calculator to do a sine regression in the form $y = a\sin(bx+c)+d$. Round all parameters to the nearest tenth. Graph along with your data to informally assess the fit of the curve. When prompted use 16 iterations always.

(c) According to this model, what is the range in temperatures the chemical reaction will include?

(d) According to this model, what is the time it takes for the reaction to complete one full cycle?

***Exercise* #8:** The maximum amount of daylight that hits a spot on Earth is a function of the day of the year. Taking $x = 0$ to be January 1st, daylight, in hours, was measured for 12 different days. The measurement was the number of possible hours of sun from sunrise to sunset.

Day	0	34	68	98	118	134	171	203	274	321	346
Daylight Hours	9.0	9.9	11.5	13.1	14.0	14.6	15.2	14.8	13.1	11.5	9.5

(a) What is the natural period of this data set?

(b) Use your calculator with the period from (a) to find an equation of the form $y = a\sin(bx+c)+d$ that fits this data, then examine the graph of the equation on the scatter plot. How good is the fit?

(c) What is the maximum amount of daylight hours predicted by the model? Show your calculation.

HOMEWORK

FLUENCY

1. Which of the following linear equations would best fit the data set shown below?

   (1) $y = 2.4x + 18.7$ (3) $y = -1.6x + 27.2$

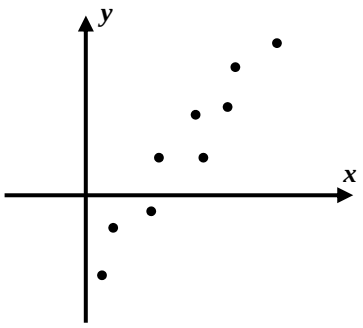
   (2) $y = -0.8x + 18.1$ (4) $y = 1.9x - 15.6$

x	2	5	9	15
y	26	17	12	4

2. A scatter plot is shown below. Which of the following *could* be the equation of the best fit line for the data set?

   (1) $y = 1.8x - 3.2$ (3) $y = -2.9x + 8.3$

   (2) $y = -3.5x - 12.4$ (4) $y = 6.5x + 3.9$



3. A line of best fit was created for a data set that only included values of x on the interval $12 \le x \le 52$. For which of the following values of x would using this model represent extrapolation?

   (1) $x = 26$ (3) $x = 14$

   (2) $x = 50$ (4) $x = 6$

4. Which of the following is true about the line of best fit for the data set given in roster form below?

   (1) It has a positive slope and negative y-intercept.
   (2) It has both a positive slope and y-intercept.
   (3) It has both a negative slope and y-intercept.
   (4) It has a negative slope and positive y-intercept.

   $\{(0,-3),(2,4),(6,10),(15,12)\}$

APPLICATIONS

5. An agronomist is studying the height of a corn plants as a function of the number of days since the corn germinated (appeared above the ground). Based on the following data, use your calculator to determine the best fit line in $y = ax + b$ form. Round all coefficients to the nearest *tenth*.

Time, x (days)	3	8	12	20	28	32	40
Height, y (inches)	2.5	4.5	6.2	9.3	12.9	14.4	16.8

6. Heavier cars typically get worse gas mileage (their miles per gallon) than lighter cars. The table below gives the weight versus the highway gas mileage for seven vehicles.

Vehicle Weight (thousands of pounds)	2.5	2.9	3.1	3.0	4.2	6.6	3.4
Gas Mileage (miles per gallon)	34	36	31	29	23	12	26

(a) Determine the best fit linear equation, in $y = ax + b$ form, for this data set. Round all coefficients to the nearest tenth.

(b) Using your model from part (a), determine the gas mileage, to the nearest mile per gallon, for a vehicle that weighs 3500 pounds.

(c) Is the prediction you made in (b) and example of interpolation or extrapolation? Explain.

(d) What is the value of the correlation coefficient to the nearest *hundredth*? Why is it negative?

7. The superintendent of the Clarksville Central School District is attempting to predict the growth in student population in the coming years. The table below gives the population for her district for selected years.

Year	1990	1992	1995	1997	2002	2005
District Population	3520	3605	3771	3860	4135	4285

(a) Find the equation for the line of best fit, in $y = ax + b$ form, where x represents the years *since* 1990 and y represents the district's population. Round all coefficients to the nearest *hundredth*.

(b) Use your model from part (a) to predict the district's population in the year 2020. Round your answer to the nearest whole number.

(c) What are the units of the slope of this linear model?

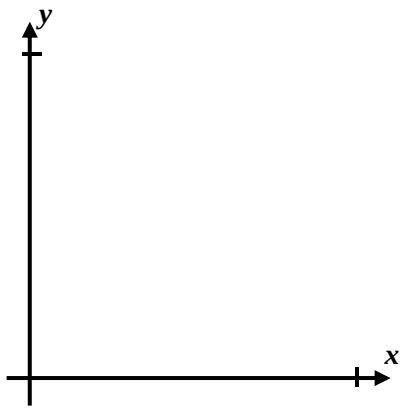
(d) What does the slope of this model represent? Think about your answer to part (c).

8. Rabbits were accidently introduced to an island where their population is growing rapidly. Biologists studying the rabbits have periodically recorded their population since they were introduced to the island. The data they took is shown below.

Years Since Introduction, x	2	5	7	11	15
Population of Rabbits, y	75	100	112	205	290

(a) Determine an exponential regression equation, in the form $y = a \cdot b^x$, that models this data. Round a to the *tenth* and b to the *hundredth*.

(b) Sketch a graph of the rabbit population below on the axes provided for $0 \le x \le 20$. Label your graphing window and your y-intercept.



(c) Based on your model in part (a), by what percent is the rabbit population growing each year?

(d) Graphically determine, to the nearest *tenth* of a year, when the rabbit population will reach 350.

9. The infiltration rate of a soil is the number of inches or water per hour it can absorb. Hydrologists studied one particular soil and found its infiltration rate decreases exponentially as a rainfall continues.

Time, t (hours)	0	1.5	3.0	4.5	6.0
Infiltration Rate, I (inches per hour)	5.3	3.1	2.4	1.6	0.7

Create an exponential model that best fits this data set. Round parameters to the nearest *hundredth*. Use your model to algebraically determine the time until the rate reaches 0.25 inches per hour. Round your answer to the nearest *tenth* of an hour. Use a logarithm in the process of your algebraic solution.

10. The soil's temperature beneath the ground varies in a periodic manner. A temperature probe was left 3 feet underground and recorded the temperature as a function of the number of days since January 1st ($x = 0$). The temperatures for 14 days throughout the year are shown below.

Day	5	36	57	94	127	153	192
Temp ($^\circ$F)	41	37	36	40	48	64	68
Day	226	241	262	289	305	337	356
Temp ($^\circ$F)	66	61	58	49	44	42	40

(a) Find a best fit sinusoidal function for this data set in the form $y = a\sin\left(bx + c\right) + d$. Round all parameters to the nearest *hundredth*. Recall that some calculators require that you input the period on this correlation (365 days).

(b) Based on your model from (a) what are the highest and lowest temperature reached in the soil?

(c) What is the average soil temperature?

(d) If the root of a particular plant species will only thrive when the soil temperature is above 50 $^\circ$F , graphically determine the interval of days over which the plant will thrive.

11. The rise and fall of the tides at a beach is recorded at regular intervals. Their period is almost 24 hours, but not exactly. The depth of a tidal marsh was measured over 3-hour time interval and the data is shown below.

Hours (since midnight)	0	3	6	9	12	15	18	21	24
Depth (ft)	5.5	8.0	10.5	11.7	10.8	8.4	5.8	4.3	4.9

Find a sinusoidal model for this data using your calculator. Place it in $y = a\sin\left(bx + c\right) + d$ form. Round all coefficients to the nearest *thousandth* (3 decimal places).

According to your model, what is the period of the tides in hours? Recall that $b \cdot P = 2\pi$.